Csapó Benő – Molnár Gyöngyvér – R. Tóth Krisztina
SZTE, BTK, Neveléstudományi Intézet

A papíralapú tesztektől a számítógépes adaptív tesztelésig

A pedagógiai mérés-értékelés technikájának fejlődési tendenciái

A tanítás és tanulás kutatásának egyik leggyorsabban fejlődő területe a mérés-értékelés. A mérések eszközei a pedagógiai tesztek, amelyek a vizsgált területről skálán kifejezhető, kvantitatív információt szolgáltatnak. Számos előnyös, mással nem helyettesíthető tulajdonságuknak köszönhetően a tesztek használata gyorsan terjed, azonban a széles körű alkalmazás felszínre hozza az egyes tesztelési technikák korlátait is. A fokozódó igények egyre újabb adatfelvételi és adatelemzési megoldások kidolgozását vonják maguk után. A tömeges felmérésre ma alkalmazható legfejlettebb technika az online adaptív tesztelés.

Bár az adaptív tesztelés alapelveit több évtizede alkalmazzák, következetes gyakorlati megvalósítását a számítógép használata tette lehetővé, ezért szélesebb körű kipróbálására is csak az utóbbi években kerülhetett sor. A számítógép alkalmazása nemcsak leegyszerűsíti a tesztelés folyamatát, hanem olyan hatékony módszereket is lehetővé tesz, amelyeket a hagyományos mérésekkel meg sem lehet közelíteni. Ugyanakkor a számítógépes tesztelés pedagógiai alkalmazása további kérdéseket vet fel, amelyekre megnyugtató választ kell találni, mielőtt a szélesebb körű elterjesztésre sor kerülne.

Tekintettel a számítógépes tesztelés kimeríthetetlen lehetőségeire, kétségtelen, hogy belátható időn belül ki fogja szorítani a papíralapú tesztelést. Iskolai kontextusban azonban csak fokozatosan lehet áttérni egy ilyen rendszerre, minden lépésben gondosan ellenőrizve, és kiszűrve a nemkívánatos mellékhatásokat. Ebben a tanulmányban áttekintjük a számítógépes tesztelés fő formáit, és bemutatjuk az adaptív tesztelés fontosabb lehetőségeit. Sorra vesszük azokat a problémákat is, amelyeket a pedagógiai alkalmazások felvetnek, és felvázoljuk a megoldás érdekében elvégzendő vizsgálatokat. A számítógépes tesztelés rövid történetére tekintettel a hatásvizsgálatok csak a közelmúltban kezdődtek el, és viszonylag kevés általánosítható eredmény áll rendelkezésre.

A kötött formátumú papír-ceruza tesztek és alkalmazási lehetőségeik

A számítógépes tesztelés sajátosságainak ismertetése előtt össze kell foglalnunk a hagyományos tesztek jellemzőit, ugyanis ezekhez viszonyítva lehet megmutatni azokat az új lehetőségeket, amelyeket a számítógépes tesztelés kínál, és hasonlóképpen így lehet megérteni azokat a problémákat is, amelyeket az új mérési technikák felvetnek. A közismert tesztek, amelyeket gyakran papír-ceruza – angol elnevezéssel Paper and Pencil, rövidítve PP *(1)* – teszteknek neveznek, nagyon fontos szerepet játszottak és játszanak ma is a tanítási-tanulási folyamatok irányításában, az oktatás eredményességének felmérésé-

ben. Ezek a tesztek többnyire rögzített formátumúak (Fixed Form – FF), ami azt jelenti, hogy a tesztek feladatait mindig azonos formai elrendezésben kapják meg a tesztelt személyek. Szigorú értelemben csak így biztosítható a teszt objektivitása, azaz hogy az mindig mindenkit egyformán mér. A tapasztalat szerint ugyanis a feladatok sorrendjének szerepe lehet a megoldás valószínűségében.

A PP FF tesztekben sokféle item (a legkisebb, önállóan értékelhető egység) fordulhat elő, változatos item-formátumokat használhatnak, ezek csoportosításának egyik dimenziója a zárt-nyitott kérdéstechnika. A zárt vagy feleletválasztós kérdések esetében előre megadott válaszokból választva kell a tesztet megoldani. Az ilyen feladatokból álló teszteket gyakran nevezik objektív teszteknek, mivel azok értékelése nem igényel személyes emberi döntéseket. A leggyakrabban alkalmazott objektív item-formátumok a többszörös választás (multiple-choice), valamint a dichotóm választás (alternatív választás, tekinthető a többszörös választás speciális esetének), amelynek egyik formája az igaz-hamis döntés (true-false). Ugyancsak objektív item-formátum az illesztés (párosítás, matching), melynek során két halmaz elemei között kell megfeleltetést létrehozni.

Minél nagyobb egy kötött formátumú teszt tétje, annál nehezebb azt kipróbálni, fejleszteni, javítani. Ez azonban nem adhat felmentést arra, hogy tömegével alkalmazzanak fiatalok sorsát eldöntő, ugyanakkor megkérdőjelezhető minőségű teszteket. A kipróbálásnak ebben az esetben is meg lehet találni a módszereit, bár azok nyilvánvalóan költségesek.

A nyitott vagy feleletalkotó (Constructed Response, CR) kérdések esetében a tesztelt személy maga alkotja meg a választ, és ennek értékelése, a válasz helyességének megállapítása további, többnyire személyes kódolói döntést igényel. A CR itemek az objektivitás szempontjából szélesebb spektrumot alkotnak a rövid választól (egy kifejezés, egy szó vagy egy szám a válasz) az esszé jellegű kérdésekig. Attól függően, hogy mennyire sokféle lehet a válasz, az értékelő (kódoló) lehetőségei is bővülnek. Így már csak bizonyos közelítéssel biztosítható, hogy egymástól független értékelők ugyanolyan módon döntsenek egy válasz helyességét illetően. A CR tesztek objektivitását az egyértelmű javítókulccsal, kódolási utasítással és az értékelők képzésével lehet javítani.

A zárt és a nyitott tesztfeladatok alkalmazása közötti választás során két ellentétes szempontot kell mérlegelni. Egyrészt az objektív itemek – mivel nem igényelnek további emberi értékelő beavatkozást – olcsóbbak, gyorsabban lehet az eredményekhez jutni. Megválaszolásuk a teszt megoldójától is kevesebb időt igényel, a kész válaszok közötti döntés gyorsabb lehet, mint a válasz önálló megalkotása. Éppen ebből következően másfajta gondolkodást igényel(het)nek, mint az önálló válaszadás, ezért esetleg csak a tudás bizonyos komponenseinek mérésére alkalmasak. A CR itemek – ha azok kódolása emberi munkával történik – kevésbé objektívek, feldolgozásuk drágább és lassúbb, viszont a tudás változatosabb formáinak felmérésére alkalmasak.

A PP FF tesztek készítésének és fejlesztésének alapjául hosszú időn keresztül a klaszszikus tesztelmélet szolgált (bővebben lásd pl. *Csapó*, 2000). Ez egy szigorú, axiomatikus matematikai elmélet, amelynek következtetései alkalmasak a tesztek minőségének jellemzésére. Az elmélet alapvető feltevése szerint minden felmért személy rendelkezik a vizsgált tulajdonság egy V valódi értékével, és minden mérés szolgáltat róla egy M mért értéket. A két érték közötti különbség a hiba, korrelációjuk pedig a teszt megbízhatóságát, reliabilitását jellemző mutató. Mivel a V közvetlenül soha nem határozható meg, az említett korrelációt sem lehet közvetlenül kiszámítani. A klasszikus tesztelmélet tételeit felhasználva azonban bizonyos mérhető adatokból lehet arra becslést adni. Például a

megismételt tesztelés adataiból, vagy a teszt belső konzisztenciájából (az itemek közötti korrelációkból). Az egyes itemek minőségét is a teszthez képest lehet megítélni: más itemekkel, főleg pedig a teszt-összpontszámmal való korreláció jól megmutatja, illik-e egy item a képbe, ugyanazt méri-e, mint a többi.

A tesztek elemzésének, a hibás, rosszul mérő itemek kiszűrésének, az itemek fejlesztésének a klasszikus tesztelméletre épülő kifinomult technikái alakultak ki, és az egymást követő kipróbálás és javítás után nagyon jó minőségű teszteket lehet készíteni. A fejlesztés eredményeként matematikailag akkor nő a reliabilitás, ha a teszt homogén, egymással magasan korreláló és közepes nehézségű itemekből áll. Ez az oktatási alkalmazások szempontjából nem mindig előnyös, mert fontos mérendő tartalmak szorulhatnak így ki a tesztből. A közepes nehézség pedig azzal járhat, hogy az átlagostól felfelé vagy lefelé eltérő teljesítmények mérésére a teszt kevésbé alkalmas.

A PP tesztek felbontása, azaz hogy egymáshoz mennyire közel álló teljesítményeket lehet velük megkülönböztetni, meglehetősen korlátozott. Ha például egy teszt 20 itemből áll és minden egyes item megoldásával 0 vagy 1 pontot lehet elérni, akkor az egymástól 5 százalék távolságra levő teljesítményeket lehet csak az adott teszttel megkülönböztetni. A felbontást az itemek (elméleti vagy tapasztalati) súlyozásával lehet finomítani, azonban a kötött formátum mellett, ha mindenki ugyanazokat a feladatokat oldja meg, a felbontás javításának komoly korlátai vannak.

A PP FF tesztekkel az említett korlátokból fakadóan csak egy viszonylag szűk képességtartományt lehet jól felmérni. Ha a teszt egy szélesebb képességtartományt fog át, akkor minden egyes felmért személynek csak a feladatok egy szűkebb sávja jelent valódi kihívást, amely a saját képességéhez közel álló feladatokat tartalmaz. A feladatok nagyobb része viszont vagy túlságosan könnyű, ezért unalmas, vagy túl nehéz, ezért frusztráló hatású lehet. Egy-egy alkalommal elvégzett tesztelésnél ezek a hatások nem túl jelentősek, ha azonban az oktatási folyamatba rendszeres tesztelés épül be, az említett negatívumok már komolyan veszélyeztetik az érdeklődést, a teszteléssel kapcsolatos attitűdöt és a feladatok megoldásához szükséges motivációt.

A teszteket az oktatásban két fő értékelési célra lehet használni, és ez a tesztekkel szemben különböző követelményeket támaszt. A formatív (segítő-formáló, fejlesztő, diagnosztikus) értékelés során a cél a tanuló közvetlen segítése, annak feltárása, mi az, amit tud, és mit kell még megtanulnia. Ebben az esetben a tanulónak érdeke az értékelővel való együttműködés, hiszen a hiányosságok kiderítése nyomán további segítséget kaphat. A formatív értékelés akkor hatékony, ha gyakori és konkrét. A szummatív (összegző-lezáró, minősítő) értékelés egy hosszabb tanulási folyamat eredményét méri. Ebből következően már nem terjedhet ki minden tudáselemre, legfeljebb mintát vehet a felmérendő teljes tudásból. Ebben az esetben felmerül a kérdés, mennyire jó ez a mintavétel, ami különösen akkor problematikus, ha maga a teljes felmérendő tudás is csak nehezen írható le.

Az előző dimenzióval szoros kapcsolatban van a tesztek alkalmazásának egy további jellemzője, az, hogy mekkora tétje van a teszteredménynek a felmért egyén számára. Ebből a szempontból megkülönböztethetjük az alacsony téttel (low stakes) és a magas téttel (high stakes) megoldott teszteket. Ez tehát nem magának a tesztnek, hanem a tesztelés kontextusának a jellemzője. Például az érettségi vizsgának kifejezetten magas a tétje, de a próbaérettségineknek elhanyagolható. Természetesen az alacsony vagy magas tét csak a két végpont megnevezése, hiszen a tét nagyságát tekintve itt is egy folytonos változóról van szó. Mindez alapvetően befolyásolja a tesztmegoldók motivációját, érdekeltségét és késztetését a mérés céljaitól idegen módszerek és eszközök alkalmazására. Például a tesztmegoldások betanulása, tiltott segédeszközök használata annál valószínűbb, minél nagyobb a tesztelés tétje. A teszt alkalmazóinak ezzel arányos erőfeszítéseket kell tenniük a tesztelés objektivitásának biztosítása, például a feladatok titokban tartása érdekében.

Ez utóbbi szempontok úgy függenek össze a tesztek formátumával és minőségével, hogy a teszteket – az előbb említett reliabilitási problémák miatt is – többszörösen ki kell próbálni, a nem jól mérő itemeket szükség esetén korrigálni kell. Amíg azonban a formatív teszteket nyilvánosan lehet kezelni, folyamatosan lehet fejleszteni és alkalmazni, a magas tétel bíró kontextusban alkalmazott kötött formátumú teszteket titkosan kell kezelni, és többnyire csak egyszer lehet alkalmazni. Ebből következik az a paradox sajátosság, hogy minél nagyobb egy kötött formátumú teszt tétje, annál nehezebb azt kipróbálni, fejleszteni, javítani. Ez azonban nem adhat felmentést arra, hogy tömegével alkalmazzanak fiatalok sorsát eldöntő, ugyanakkor megkérdőjelezhető minőségű teszteket. A kipróbálásnak ebben az esetben is meg lehet találni a módszereit, bár azok nyilvánvalóan költségesek.

A kötetlen formátum és a valószínűségi tesztelmélet lehetőségei

Az oktatási kontextusban alkalmazott mérések többnyire nem egyetlen kötött formátumú tesztet igényelnek, mert például olyan nagy tudásterületet vizsgálnak, vagy olyan széles képességfejlődési spektrumot kellene átfogniuk, amelyek technikai okokból sem férnek bele egyetlen tesztbe. A probléma megoldására számos technika született. Ezek közé tartozik a teljes lefedés elve, amikor egy nagyobb tudásterület teljes felméréséhez a lehetséges összes feladat elkészül. Ilyen megoldást dolgozott ki Nagy József az általa irányított program elméleti keretéül, amikor a fontosabb iskolai tárgyak teljes tudásanyagát magában foglaló tesztek készültek (*Nagy*, 1972). Ilyen esetben az elkészült feladatokat ekvivalens tesztváltozatokba sorolják úgy, hogy minden egyes tesztváltozat kezelhető méretű legyen. Így, bár az országos reprezentatív felmérések során egy tanuló mindig csak az összes feladat egy részét oldotta meg, a felmérés egészéből az összes tudáselem elsajátításáról képet lehetett alkotni.

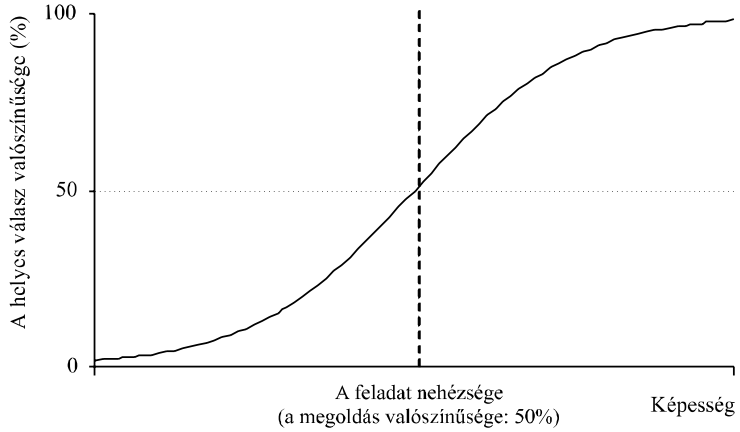
Egy másik megoldás a feladatbankok alkalmazása, amikor lényegében a teljes lefedés előzőekben bemutatott elveit alkalmazva, tesztváltozatokba sorolva kerül sor a feladatok bemérésére. Ezután az összes feladat egy feladatbankot alkot, amelyből a konkrét felmérések igényeinek megfelelően lehet kiválasztással vagy véletlen sorsolással a konkrét felmérések céljaira teszteket összeállítani. Erre a megoldásra is lehet egy korai példát bemutatni a magyarországi gyakorlatból (*Nagy*, 1976).

További probléma – különösen a képességtesztek esetében –, hogy a tanulók között nagyobbak a különbségek, mint amekkorát egy kötött formátumú teszttel le lehet képezni. Ha a teszt túl széles spektrumot próbál átfogni, minden tanuló csak néhány olyan feladatot talál, amelyik tudásszintjéhez közel áll, a feladatok többsége pedig vagy túl könnyű, vagy túl nehéz. Ha a tanulók a feladatokból egyénileg a képességszintjükhöz közeli válogatást kapnak, pontosabban be lehet határolni a konkrét fejlettséget.

A klasszikus tesztelmélet által kínált eljárásokat alkalmazva ki lehet számítani a teszt sokféle jellemzőjét, azonban a paraméterek többsége szigorúan véve csak a teszt bemérésére alkalmazott minta (tanulócsoport) esetében lesz érvényes. A már korábban említett, valamint további, itt nem elemzett problémák megoldására a klasszikus tesztelmélet kereteit továbbfejlesztve, illetve a PP tesztek kötött formátumát megbontva számos előremutató megoldás született. Azt a problémát azonban, hogy miként lehet feladatokhoz különböző paramétereket, mindenekelőtt a nehézséget jellemző mértéket rendelni, függetlenül attól, hogy éppen melyik tesztben alkalmazzuk, a valószínűségi tesztelmélet (más neveken: modern tesztelmélet, Rasch-modell, Item Response Theory, IRT) oldotta meg. Ezzel megnyílt az út a változatos összetételű, kötetlen formátumú tesztek alkalmazása előtt.

A valószínűségi tesztelmélet a mérés során elkövetett hibát és az itemek tulajdonságait más módon, nem determinisztikusan, hanem valószínűségi alapon kezeli. A valószínűségi tesztelméleti modellek közül speciális tulajdonságai miatt, amelyek lehetővé teszik a mintafüggetlen, illetve tesztfüggetlen értékelést (két személy összehasonlítása függet-

len attól, hogy melyik itemen tesszük azt, illetve két item összehasonlítása független attól, hogy milyen képességszintű személy oldotta meg azokat, részletesebben lásd *Molnár*, 2006), kiemelt figyelmet fordítunk a dichotóm Rasch-modellre (a nem dichotóm modellekről részletesebben lásd: *Molnár*, 2008). A Rasch-modell az itemek paraméterezése és a személyek képességszintjének meghatározása során abból az egyszerű gondolatból indul ki, hogy a magasabb képességszintű személy nagyobb valószínűséggel oldja meg ugyanazt az itemet, mint az alacsonyabb képességszintű, illetve egy item akkor nehezebb, ha azt kisebb valószínűség mellett oldják meg, mint a másikat (*Rasch*, 1960; idézi *Griffin*, 1999). Ennek megfelelően minden egyes itemhez hozzárendel egy itemkarakterisztikus görbét, ami alapján megállapítható, hogy az egyes képességszintű diákok milyen valószínűség mellett válaszolnak jól az adott itemre. A magas képességű diák jó válaszának valószínűsége közel áll a 100 százalékhoz, míg az alacsony képességszintű diáké a 0 százalékhoz. Egy átlagos nehézségű feladat esetén az átlagos képességszintű diák helyes válaszának valószínűsége 50 százalék (1. ábra), mivel az item nehézségi indexe azon személy képességparamétere alapján definiált, aki 50 százalék valószínűség mellett oldja meg jól az adott feladatot.



1. ábra. Egy példa az itemkarakterisztikus görbére

Miután az itemek nehézségi indexei a diákok képességszintjei alapján definiáltak, ezért az itemek nehézségét és a diákok képességszintjét közös képességskálán tudjuk ábrázolni. A Rasch-modell speciális objektivitása (teszt- és mintafüggetlensége) miatt, ha ismerjük egy diák képességszintjét, meg tudjuk mondani, hogy milyen valószínűséggel oldana meg egy olyan itemet, amelynek nehézségi indexe értelmezhető a közös képességskálán, anélkül hogy a diáknak a valóságban meg kellene oldani azt (mintafüggetlenség). Megfordítva, a közös képességskálán lévő itemekből válogatott teszt alapján (tesztfüggetlenség) bármely diákhoz hozzá tudjuk rendelni képességparaméterét anélkül hogy az összes feladatot, itemet meg kellene oldania. Ehhez viszont az itemeket közös képességskálán kell jellemeznünk. Ezt a problémát horgony-itemek alkalmazásával hidalhatjuk át.

Horgony-itemeknek nevezzük a különböző tesztek azonos, átfedő feladatait. Ezen horgony-itemek segítségével a meglévő itemekhez hozzáskálázhatók az újonnan felvett feladatok. Miután számos azonos tulajdonságot mérő itemet paramétereztünk ezen a módon, felépíthető belőlük egy feladatbank, ami a hatékony tesztelés alapját képezi.

Egy jól felépített feladatbank minőségét négy faktor segítségével lehet jól jellemezni.

(1) A feladatbank nagysága, azaz a feladatbankban szereplő itemek száma. Minél kevesebb itemből áll egy feladatbank, annál nagyobb annak valószínűsége, hogy bizonyos

itemek gyakrabban előfordulnak, azaz könnyebben megjegyezhetővé válnak. Ennek hatására romlana a teszt validitása. Ezt kiküszöbölhetjük úgy, hogy több száz (minimum 300) feladatból (*Weiss*, 2004; *Van der Linden, Ariel és Verdkamp*, 2006) állítjuk össze a feladatbankot, illetve a tesztelést irányító algoritmus szabályrendszerét úgy alakítjuk ki, hogy a program az adott személyre jellemző leginformatívabb öt item közül véletlenszerűen válasszon egyet.

(2) Az itemek homogenitása, azaz a valószínűségi számításokhoz alapul vett matematikai modellhez való illeszkedése. Ez azt jellemzi, hogy mennyire azonos az itemek diszkrimináló ereje (erről részletesen lásd *Molnár*, 2006).

(3) Az itemek diszkrimináló ereje. Minél nagyobb diszkrimináló erővel rendelkező itemeket kell használni, mégpedig úgy, hogy azok átlagos nehézségi szintje lefedje a teljes képességtartományt. Egy adott item azon a képességszinten differenciál legjobban, ami azonos nehézségi paraméterével. A többi képességtartomány lefedésére más nehézségi indexű jól diszkrimináló itemek alkalmazása hatékony.

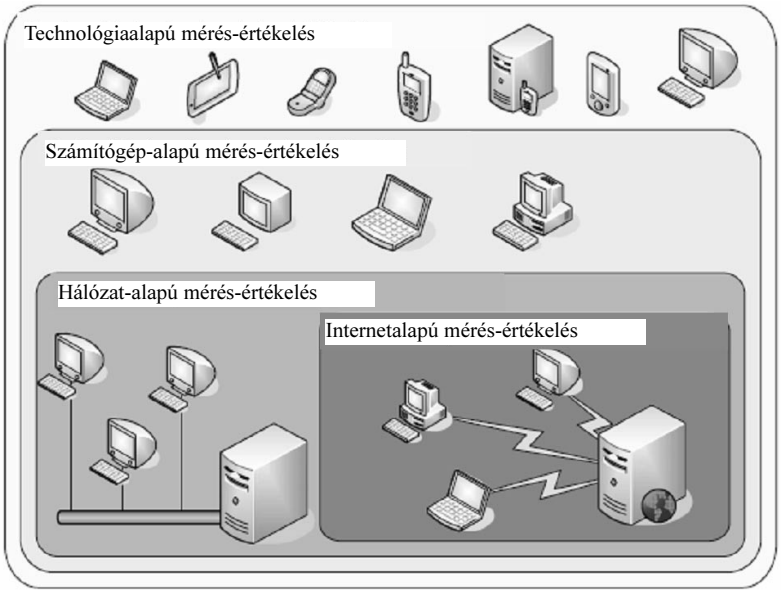
(4) Az itembank validitása. Az itemek ugyanazt a tulajdonságot, ismertetőjegyet, képességet, készséget mérik, amelyet a tesztelés elméleti keretei rögzítenek. Emellett a megfelelő feladatszám biztosítja, hogy ne lehessen a megoldásokat formai elemek alapján előre betanulni, ne lehessen magára a tesztelésre „edzeni” (test coaching) a tesztelendő képesség valódi elsajátítása nélkül.

A számítógépes tesztelés

Lényegében a számítógép oktatási célú alkalmazásával egy időben megjelent a számítógépes tesztelés. A feleletválasztós feladatokat minden nehézség nélkül át lehetett ültetni számítógépre, és ahogy a számítógépek fejlődtek, úgy alakultak ki az egyre fejlettebb számítógépes technikák. A számítógép-alapú tesztelés (Computer Based Assessment – CBA) általában minden komputeres értékelést magába foglal; kicsit tágabb értelemben használják még a technológiaalapú tesztelés (Technology Based Assessment – TBA), illetve az elektronikus tesztelés (e-Testing) kifejezéseket is. Az alkalmazott technológia szerint megkülönböztetett szintek egymásra és egymásba épülését a 2. ábra szemlélteti.

A technológiaalapú mérés magába foglalja az összes olyan mérési-értékelési rendszer alkalmazását, ahol az adatgyűjtésre valamilyen információs-kommunikációs technológiai eszközt használunk. Annak ellenére, hogy ez az eszköz általában a számítógép, mégis a számítógépes mérés-értékelés halmazát magába foglaló bővebb halmazként megkülönböztetjük ezt a kategóriát. Ennek oka, hogy bizonyos esetekben a közvetítő eszköz nem feltétlen a számítógép: lehet PDA, mobiltelefon, szavazórendszer stb. (ezek iskolai alkalmazásáról lásd *Molnár*, 2007), amelyek egy része alkalmas arra, hogy a nap bármely időszakában bizonyos kérdéseket tegyen fel a mérésben résztvevőnek – attól függetlenül, hogy az illető helyileg hol van –, aki arra azonnal válaszolni tud.

A technológiaalapú mérésen belül természetesen a legtöbb lehetőséget a számítógép-alapú értékelés kínálja, ennek alkalmazása ma minden másnál sokkal elterjedtebb. A számítógép-alapú mérés-értékelés során az alkalmazott teszt a számítógép monitorán jelenik meg (on-screen presentation), a tesztelt személy pedig szintén a számítógép segítségével (billentyűzet, egér stb.) adja meg válaszát. A válaszok rögtön elektronikusan rögzítésre kerülnek, majd a válaszok elemzése is általában a számítógép felhasználásával történik. A számítógép-alapú tesztelésbe beletartozik annak mind hálózati, mind interneten keresztül történő alkalmazása. Ha semmilyen hálózatot (helyi hálózat, internet) nem vonunk be a tesztelés lefolytatásába, akkor a tesztelést végző programot, feladatlapot minden egyes számítógépre installálni kell. Az esetleges változtatásokat minden egyes számítógépen külön regisztrálni kell, majd az adatokat minden egyes számítógépről be kell gyűjteni.



2. ábra. A technológiaalapú, a számítógépalapú, a hálózat- és internetalapú mérés-értékelés hierarchikus viszonya (Jurecka és Hartig, 2007 alapján)

A hálózatalapú mérés-értékelés a számítógépes tesztelés egy olyan alkalmazását jelenti, amikor a teszt, a feladatok, a tesztelést végző program egy adott számítógépes hálózaton belül érhető csak el. Ez a hálózat lehet helyi (LAN), vagy az internet, vagy a kettő kombinációja (*Jurecka és Hartig*, 2007). A hálózatalapú mérés egy gyakori alkalmazása, amikor az adott hálózaton belül egyszerre több gépen zajlik a tesztelés, azt egy külön számítógépről irányítják, ahol az adatok összegyűjtése, elemzése történik. A tesztelés előtt minden egyes adatfelvételben részt vevő gépre felinstallálják a szükséges szoftvert. A kiértékelés szoftvertől függően vagy a helyi számítógépen, vagy a központi szerveren történik.

Az internetalapú tesztelés során az adatfelvétel kizárólagosan az interneten keresztül történik. Az adatfelvételben részt vevő személynek csak internetkapcsolatra és egy internetes böngészőre van szüksége a tesztelésben való részvételhez. Ebben az esetben nincs szükség arra, hogy a helyi számítógépen fusson a tesztelő program. A vizsgázó azonosítójával be tud lépni a rendszerbe, ahol csatlakozik a tesztelő szoftverhez, ami a szerverrel kommunikálva választja ki a diák számára a megoldandó feladatokat. Mind a feladatok, itemek, mind a szoftver a szerveren és nem lokálisan a számítógépen van. A válaszok, adatok tárolását és kiértékelését is a központi szerver végzi. Ebből adódóan könnyebb és gyorsabb mind az itembank módosítása, mind a szoftver frissítése. További előny, hogy ha a szoftver külső gépen fut, nem kell minden iskolának saját szoftverrel rendelkeznie.

A számítógépes tesztelésre kifejlesztett rendszereket az alkalmazott médiumon kívül egy másik dimenzió mentén is csoportosíthatjuk: a feladatlapok, feladatok, itemek típusa, személyre szabottsága mentén. Ezen változó minden egyes szintje megvalósítható a fent nevezett halmazok, részhalmazok bármelyikében. A továbbiakban e dimenzió mentén különítjük el egymástól az egyes lehetőségeket.

A számítógépes tesztelés legegyszerűbb formája (a PP teszttől való eltávolodás tekintetében a nulladik szintjének is nevezett megoldás) a PP tesztek egyszerű, az eredetivel megegyező formában való digitalizálása. Ebben az esetben csak a feladatokat közvetítő

eszköz, vagyis a médium változik meg. A feladat a papír helyett a képernyőn jelenik meg, a válaszadás billentyűvel, egérrel, érintőképernyővel vagy egyéb elektronikus eszközzel történik. A tesztelés továbbra is lineáris marad, a feladatok azonos sorrendben jelennek meg minden egyes tesztelt személy előtt. Érintőképernyőt használva a PP teszteléssel való egészen közeli hasonlóságot lehet elérni, a vizsgázó – az érintőképernyő technológiájának függvényében – egy digitalizáló vagy egy közönséges toll segítségével jelöli meg válaszát. Egér vagy billentyű használata esetében már szükség van némi technikai készségre, ha pedig a billentyűzettel hosszabb szövegeket kell bevinni, már számíthat a gépírási készségek fejlettsége is. A legtöbb létező számítógép-alapú teszt ehhez hasonló formátumú, feleletválasztós feladatokból álló standardizált teszt (*Jurecka és Hartig*, 2007).

A számítógépes tesztelés már ezen a nulladik szintjén is számos előnnyel jár. Annak ellenére, hogy a tesztelt személy számára nem jelent nagy különbséget, a javítás, kódolás, rögzítés munkafázisait ki lehet iktatni, vagy jelentősen le lehet egyszerűsíteni. Objektív feladattechnikát alkalmazva a teszt kiértékelése azonnal megtörténik, az eredmény rögtön rendelkezésre áll. A PP tesztelés során emberi munkára van szükség a válaszok javításához, rögzítéséhez, ami magában foglalja az adatvesztés lehetőségét, az adatminőség romlását is.

Lényegében a számítógép oktatási célú alkalmazásával egy időben megjelent a számítógépes tesztelés. A feleletválasztós feladatokat minden nehézség nélkül át lehetett ültetni számítógépre, és ahogy a számítógépek fejlődtek, úgy alakultak ki az egyre fejlettebb számítógépes technikák. A számítógép-alapú tesztelés (Computer Based Assessment – CBA) általában minden komputeres értékelést magába foglalt; kicsit tágabb értelemben használják még a technológiaalapú tesztelés (Technology Based Assessment – TBA), illetve az elektronikus tesztelés (e-Testing) kifejezéseket is.

Az adatminőség javulásával a mérés egyik minőségi kritériumát, egyik jóságmutatóját, az objektivitást növeljük. Az adatfelvételi objektivitás esetén a teszteredménynek függetlennek kell lennie az adatfelvevő személyétől (*Csapó*, 2000), azaz a vizsgázó teszten elért eredménye nem függhet a mérőbiztos személyétől. Ez teljes mértékben biztosított, ha a feladatokat a számítógép közvetíti, és a tesztek megoldásának környezeti feltételeit is egyszerűbben lehet egységesíteni. A számítógép nem fáradt, nem unatkozik, nem frusztrált (*Becker*, 2004), nem sürgeti a tesztbeadást, valamint megtakaríthatjuk a tesztet felvevő tanárok felkészítését is. Az adatfelvétel minőségének javításához az is hozzájárul, hogy a feleletválasztós feladatokra (mind alternatív választás, mind többszörös választás esetén) adott válaszok véletlenszerűségét minimalizálhatjuk, hiszen a diákok nem tudnak előre-hátra lapozni a feladatsorban.

A számítógépes tesztelés során növelhetjük a teszt értékelésének objektivitását, minőségét is, mivel egyrészről a diákok eredményét nem befolyásolja a javító szigorúsága, másrészről megszűnnek a javítás, kódolás és rögzítés során keletkezett kiértékelési hibák. A számítógépes kiértékelés segítségével akárhányszor lefuttatjuk a kiértékelést, mindannyiszor ugyanarra az eredményre jutunk. Az automatikus tesztkiértékelés gyors és egyszerű folyamat, még összetett kiértékelő algoritmusok esetén is. Az emberi figyelmetlenség miatt bekövetkező kiértékelési hiba az esetek 10 százalékában fordul elő (*Butcher*, 1987. 17.; idézi *Becker*, 2004). Fontos megjegyezni, hogy ha automatikusan értékelünk ki, akkor nem csak a feladat javításakor előforduló hibákat zárhatjuk ki, hanem a tradicionális tesztelés alkalmával végzett adatrögzítéskor bekövetkező elgépelések hi-

báit (ha például 45-öt rögzítenek 54 helyett) is. Az automatikus kiértékelés lehetővé teszi továbbá az egyszerű dokumentációt, szervezést, nagyobb tesztadat-mennyiségek (adatbankok) összekötését, és gyors lehívhatóságot (*Becker*, 2004) biztosít.

A számítógépes tesztelés segítségével az adatok gyorsan aktualizálhatók, valamint azonnali visszacsatolási lehetőséget nyújt a diákok, tanárok, iskola, régió stb. számára. Az azonnali visszacsatolás pedig hozzájárul az oktatási-tanulási folyamat minőségének javulásához.

A számítógép-alapú tesztelés induló költsége jelentősebb mértékű, mint egy papír-ceruza tesztelés lebonyolítása, viszont a rendszer kiépítése után a számítógép alapú tesztelés számos megtakarítási lehetőséget kínál. A számítógépes kiértékelés segítségével kiküszöbölhetjük a tesztlapok nyomtatását, fénymásolását, csomagolását, szállítását, válaszlapok készítését, stb., ezáltal az eszközköltség is jelentősen csökken. A tesztek javítására nem kell javítókat alkalmazni, a rögzítésre rögzítőket, sőt az alapstatisztikai számítások abban a pillanatban elkészülnek, ahogy a diák befejezte az utolsó item megoldását. Rose és munkatársai (1999) szerint a számítógépes teszteléssel a dokumentációs költségek 2/3-át meg lehet spórolni.

Az elektronikus rendszerre való áttérés ezen nulladik fokán már lehetőség adódik a papíralapú és a számítógép-alapú tesztelés hatékonyságának, eredményeinek összehasonlítására. A szakirodalomban számos kritikus észrevétellel is találkozunk a számítógépes teszteléssel kapcsolatban. Leggyakrabban a számítógépes tapasztalat hiányát és a számítógéptől való idegenkedést említik. Ahogy azonban az információ- és kommunikációtechnológiai (IKT) eszközök terjednek a hétköznapi életben, ennek a tényezőnek a súlya egyre kisebb lesz. Nem szabad viszont megfeledkezni arról, hogy mindaddig, amíg a számítógéphez való hozzáférés tekintetében iskolák, társadalmi csoportok és családok között jelentős különbségek lesznek, gondosan meg kell vizsgálni, nem hoz-e az alkalmazott eljárás egyeseket hátrányos helyzetbe. Gondoskodni kell arról, hogy az alkalmazott technika kezelése senkinek ne okozzon nehézséget, és ne vonja el a figyelmét az érdemi feladatmegoldó munkától. Ennek egyik legbiztosabb módja magának a számítógépes tesztelésnek az elterjesztése és gyakori alkalmazása.

A PISA 2006-os vizsgálatban már opcionálisan szerepelt a természettudományi tudás számítógépes felmérése (Computer Based Assessment of Science – CBAS), amiből kiderült, hogy a kétféle médiummal (PP és TBA) elért eredmények között komoly különbségek voltak. A PISA 2009-es felmérésben az elektronikus szövegek olvasása (Electronic Reading Assessment, ERA) *(2)* már a szövegértés terület önálló részskálája lesz (*OECD*, 2007). A következő felmérési ciklusokban a CBA mind nagyobb szerepet kap, és belátható időn belől teljesen megszűnik a PP felmérés. A PISA szakértői ettől azt várják, hogy csökken a szervezési költség és a diákok tesztelés során igénybe vett ideje is. Hosszú távon számos további előnye is lesz a számítógép-alapú tesztelés bevezetésének: lehetőség nyílik a gondolkodás olyan aspektusainak mérésére, amit papíralapú teszteléssel nem lehet megvalósítani (ez már a számítógépes tesztelés első, második és harmadik szintjén mutatkozik meg).

A számítógépes tesztelés első szintjén megtörténik a technológia adta lehetőségek további kihasználása, ezáltal gazdagíthatjuk a tesztelés során alkalmazott itemek típusát. Alkalmazhatunk multimédiás (hang, mozgókép, animáció, szimuláció, interaktív szimulácó stb.) elemekkel gazdagított itemeket is, sőt a kiegészítő technológiák alkalmazásával lehetőség nyílik a fogyatékkal élő tanulók tudásának mérésére is. A „látási, hallási és a kézírás készségével kapcsolatos problémák jó része kiküszöbölhető" (*Kárpáti*, 2002. 8.). Ezenfelül a diákok konkrét válaszán kívül további adatokat gyűjthetünk a tesztelés során a tanulókról. Mérhetjük a diákok egyes feladatok megoldásához szükséges idejét, rögzíthetjük reakcióikat, az egér mozgatását, a billentyűk lenyomása között eltelt időt, szemmozgásukat, amelyek további adatokat szolgáltatnak a figyelemre, gyorsaságra, olvasási képességre (visszaugrások száma) stb. vonatkozólag.

A számítógépes tesztelés második szintjén lehetőség nyílik egyrészt automatikus itemgenerálásra – így bizonyos típusfeladatok mindig új formában jelenhetnek meg, például a szöveges feladatokban mindig más-más számértékek szerepelnek –, másrészt az itemek előzetes csoportosítása után a létrehozott csoportokból randomizált itemválasztásra. Ezáltal biztosíthatjuk, hogy a tesztelés során mindenki azonos nehézségű, de különböző feladatokat kapjon.

A számítógépes tesztelés harmadik szintjén egy teljes mértékben parametrizált, indexelt és egy azonos nehézségi, illetve képességskálán leírható feladatbank áll a tesztelés hátterében. Ha a feladatbankból az egyes feladatok kiválasztása a vizsgázó előző válaszainak függvényében történik, adaptív tesztelésről beszélünk.

A számítógépes adaptív tesztelés

A számítógépes tesztelés igazán nagy lehetősége azonban az adaptivitás: lehetőség van arra, hogy attól függően kaphassanak a vizsgázók újabb feladatokat, miképpen oldották meg az előzőt. A számítógépes adaptív tesztelés (Computerized Adaptive Testing – CAT) a teljesítmények sokkal finomabb felbontását, mérését teszi lehetővé. Elméletileg tíz feladat megoldásával 2^{10}, azaz 1024 lehetőség közül választhatjuk ki, hogy pontosan milyen a vizsgázó képessége egy adott területen. Elméletileg, természetesen, mert a gyakorlatban ehhez az kellene, hogy legyen 1024 olyan feladat, amelyik nehézsége egyenletesen fedi le a felmérendők képességtartományát. Ilyen feladatbankot azonban szinte lehetetlen elkészíteni, mivel a feladatok pontos nehézségét csak empirikus úton lehet meghatározni, és nem lehet „rendelésre” gyártani előre meghatározott nehézségű feladatokat. Mindenesetre ez a becslés jelzi az adaptív tesztelés elméleti lehetőségeit, de egyben a megvalósítás korlátait is.

A hagyományos papír-ceruza tesztelés, illetve a tesztek digitalizált formában történő felvétele során minden egyes személy számára ugyanazon feladatok, ugyanabban a sorrendben adottak. Ezzel szemben az adaptív tesztelés során minden egyes személy másmás feladatokat, a számára leginkább diagnosztikus erővel bíró feladatokat kapja megoldásra, azaz elhanyagolható annak valószínűsége, hogy minden egyes személy ugyanazon feladatokat ugyanabban a sorrendben oldja meg. Ezáltal új lehetőségek nyílnak meg a mérés-értékelés területén.

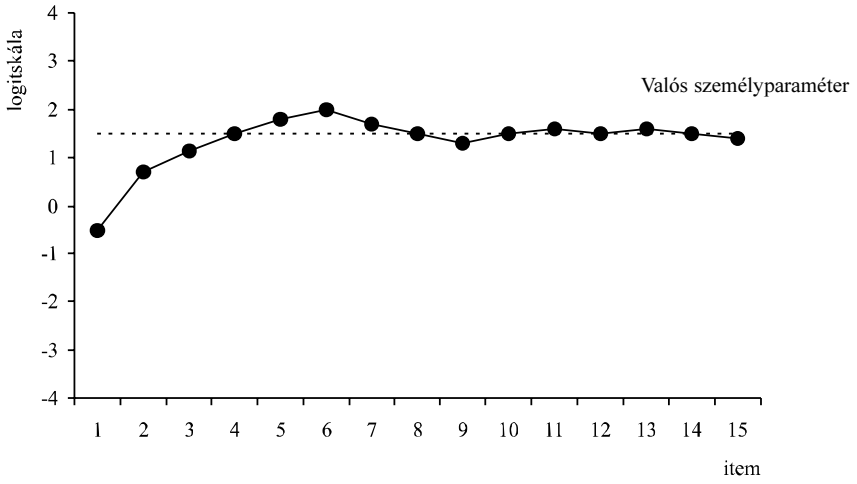
A vizsgáztatás, mérés-értékelés e formáját analógiába állíthatjuk a szóbeli vizsgáztatással, ahol a vizsgáztató a kérdéseit gyakran a vizsgázó képességeihez igazítja. Ha a vizsgázó egy közepes nehézségű kérdésre helyes választ ad, akkor a vizsgáztató következő kérdése általában egy nehezebb kérdés, míg ha helytelen a kérdésre adott válasz, akkor a közepes nehézségűnek számító kérdést egy könnyebb kérdés követi. A vizsga végén az értékelés annak függvényében történik, hogy milyen nehézségű kérdésekre tudott még helyesen válaszolni a vizsgázó. Ha csak nehéz kérdéseket fogalmazna meg a vizsgáztató, akkor az alacsonyabb képességű vizsgázók értékelése nehézkessé válna, míg csak könnyű kérdések esetén nem lehet a jobb képességű vizsgázókat differenciálni.

Az adaptív tesztelés során a fentiekhez hasonló módon történik az itemek, feladatok kiválasztása, csak a szóbeli vizsgával ellentétben néhány tényező tekintetében pontosabb, egzaktabb módon (*Frey*, 2007). A tesztelés során kiválasztásra kerülő itemeket, kérdéseket a korábban kiválasztott feladatokra adott válaszok milyensége határozza meg. Ez az eljárás azt a célt szolgálja, hogy minden egyes személy elé csak olyan itemek kerüljenek, amelyek a lehető legnagyobb információval, diagnosztikus erővel bírnak az adott személy vizsgált képességszintje tekintetében, azaz amelyek lehetőleg a legközelebb vannak valós képességszintjéhez. A legtöbb esetben ez a kiválasztás az itemek nehézsége alapján történik. A magasabb képességszintű egyének nehezebb, az alacsonyabb képességszintűek átlagosan könnyebb feladatokat kapnak a tesztelés során. Ezzel az el-

járással elkerülhető, hogy az alacsonyabb képességszintűeket esetlegesen számukra túl nehéz feladatokkal frusztráljuk, illetve a magasabb képességszintűek tesztelésre szánt idejét a könnyebb feladatok megoldásával töltsük ki. Az itemek kiválasztása egy előzetesen meghatározott algoritmus alapján történik. Ez az algoritmus egy olyan szabályrendszer, ami meghatározza az első és a rákövetkező itemek kiválasztását, továbbá specifikálja a tesztelés befejezésének kritériumait is.

Az adaptív tesztelés megvalósulását egy példán keresztül szemléltetjük. Adott 300 azonos tulajdonságot mérő dichotóm item. Minden egyes itemhez – korábbi mérések alapján – hozzárendeltük a nehézségi paraméterét. Az 1,5 logitegység képességszintű személy (ez az információ a valóságban természetesen nem áll előzetesen rendelkezésre: éppen ez az, amit keresünk) tesztelésének folyamatát mutatja a 3. ábra, ahol a szaggatott vonal a személy jelen esetben ismert képességszintjét, a fekete jelölő pedig a szimulált tesztelés során megoldásra kerülő itemek nehézségi szintjét mutatja, ami egy idő után oszcillál a személy képességparamétere körül.

Első lépésként a személy kap egy közel átlagos nehézségű ($\delta = -0{,}5$) itemet, amit jelen esetben, ismerve a tesztelt személy képességszintjét, magas valószínűséggel helyesen old meg (ennek okáról lásd *Molnár*, 2006). A vártnak megfelelően a jó megoldást egy nehezebb ($\delta = 0{,}7$) item követi (ennek a megoldási valószínűsége már alacsonyabb, de még mindig magas). Az előzetes feltételezésnek megfelelően ezt az itemet is jól oldotta meg a vizsgázó, ezért következő lépésben egy még nehezebb itemet kap ($\delta = 1{,}15$). Ez a nehézségi szint már közelíti a mért személy képességszintjét, ezért az általa adott helyes válasz valószínűsége is közeledik az 50 százalékhoz, ami akkor a helyes válasz valószínűségi szintje, ha megegyezik a személy képességparamétere az item nehézségi szintjével. Az egymást követő feladatok nehézsége egész addig növekedik, amíg a vizsgázó először helytelen választ nem ad. Ennek bekövetkezése után az előzőnél könnyebb feladatot kap megoldásra. Ha azt sem tudja megoldani, akkor egy még könnyebb feladatot kap egészen addig, amíg helyes választ nem ad. Ha ez bekövetkezett, ismét egy nehezebb item következik. Ez a folyamat egészen addig tart, amíg az előre meghatározott adaptív algoritmus szabályrendszere alapján befejezhető a tesztelés. Ez bekövetkezhet akkor, ha például (1) bizonyos, előre meghatározott mennyiségű item megoldásra került; (2) a személyparaméter becslési hibája a megengedett hibahatáron belül mozog; (3) eltelt a tesztelésre fordítható idő; (4) az itembankban előforduló összes item bemutatásra került.



3. ábra. Egy adaptív tesztelés menetének illusztrációja. A pontok az itemek nehézségi szintjét reprezentálják

A számítógépes adaptív tesztelés összességében kevesebb item használatával és rövidebb idő alatt pontosabb képességszint-meghatározást tesz lehetővé. A technológia adta lehetőségek kihasználásával növelhetjük a tesztelés során felhasznált itemek típusát például azzal, hogy alkalmazhatunk multimédiás elemekkel gazdagított itemeket is. A számítógép lehetővé teszi a gyors és hiba nélküli értékelést, visszajelentést, a kiértékelés és tesztelés folyamatában nincs szükség javításra, rögzítésre, nyomda- és postaköltségre, aminek az előnye legjobban a nagymintás vizsgálatok esetében mutatkozik meg. A teszt adaptivitásánál fogva nő a tesztbiztonság, mivel a jól és rosszul megoldott itemek, illetve az előre meghatározott algoritmus függvényében személyre szabott tesztet tölt ki mindenki, azaz megszűnik a súgás, lesés és előre kondicionált itemek problémája, viszont megmarad a standardizált mérés. Ebből adódóan gyakran ismételhető, nem szükséges minden egyes mérés során új teszteket kidolgozni, mert a rendszer az előre kifejlesztett adatbankból válogatja össze a diák képességszintjének legpontosabb meghatározásához szükséges tesztet. Ezért a rendszer alkalmas arra, hogy a tanulókat megfelelő gyakorisággal felmérje, ezáltal állandó visszajelzést biztosítson aktuális fejlettségük állapotáról.

Az azonos feladatbankon alapuló eredmények a közös nehézségi, illetve képességskálán definiált itemek miatt viszonyíthatók egymáshoz, azaz a tanuló korábbi fejlettségi szintjével összevethető az aktuális eredménye, még akkor is, ha összességében minden egyes alkalommal más itemeket oldott meg. Ezzel kiküszöbölődik a longitudinális fejlődésvizsgálatok egyik alapproblémája, miszerint ugyanazt a tulajdonságot többször egymás után ugyanazzal a teszttel kell felmérni, azonban így a tesztfeladatok egyre ismerősebbek lesznek, ami torzíthatja az eredményeket.

A teszt eredménye összevethető a többi diák azonos mérésben megoldott eredményével, illetve az adatbank felépítése és az adott képességterület skálázása során meghatározott, tudományosan kidolgozott standardokkal. Ennek következtében a papíralapú keresztmetszeti vizsgálatok lebonyolítására könnyen megvalósítható a standardizált longitudinális vizsgálat.

A CAT lényegében személyre szólóvá teszi a mérést azáltal, hogy minden tanuló többségében a saját képességszintjének megfelelő feladatokat old meg. Ezáltal a mérés egésze sokkal szélesebb képességsávot tud átfogni, mint a PP FF tesztek, mégis minden egyes esetben érzékenyebb, azaz az FF teszteknél kisebb különbségeket ki tud mutatni. A képességszinthez közel eső feladatok minden diák számára optimális kihívást jelentenek, így a munka nem válik unalmassá, és nem okoz túlzott szorongást sem. A tesztelési folyamat az optimális tapasztalatok (a flow-élmény, lásd *Csíkszentmihályi*, 1997) sávjában marad. Mindez előnyösen hat az érdeklődésre és a motivációra, aminek a tesztek gyakori alkalmazásánál meghatározó jelentősége van.

A felsorolt előnyös tulajdonságok nagyon vonzóvá teszik a CAT alkalmazását, azonban egy jól működő CAT rendszer kidolgozása rendkívül bonyolult feladat. Még abban az esetben is, ha a mérendő tulajdonság egyszerűen leírható, a feladatok empirikus nehézségét csak megfelelő mintán való kipróbálással lehet meghatározni. Az elkészült feladatok jelentős részéről már az első kipróbálás során kiderül, hogy valamilyen szempontból hibásak, nem differenciálnak, nem illeszkednek a modellbe stb. A szűrőn átjutó feladatoknak pedig éppen ezért nem megfelelően szóródik a nehézsége a felmérendő spektrumon. A fejlesztés újabb fordulóiban további feladatok készülnek, már szándékoltan könnyebbek vagy nehezebbek a még „üres” képességtartományok lefedésére. Egy feladat elkészítése során a nehézségével „beletalálni” egy adott képességtartományba szinte lehetetlen, ezért általában többtucatnyi feladatot el kell készíteni, ki kell próbálni, mire közülük legalább egy megfelel az elvárásoknak. Nehezíti az elvégzendő fejlesztő munkát, ha mindezt iskolai kontextusban kell elvégezni, hiszen így bizonyos tudást csak a tanév megfelelő szakaszában lehet felmérni, így korrekciós fejlesztő ciklusokra esetleg csak egy újabb év múlva kerülhet sor.

Perspektívák és problémák

Mint minden új, a hagyományostól eltérő módszer bevezetésekor, a számítógépes tesztelés esetében sem csupán a lehetőségekre, hanem a problémák és veszélyek elemzésére is figyelmet kell fordítani.

A számítógépes tesztelés megvalósításának egyik alapfeltétele a megfelelő hardver- és szoftverkörnyezet megteremtése. A technikai feltételek megteremthetőségének kérdése egyrészről az iskolákban, másrészről a tesztelés központjában merül fel. Az iskolákban a csoportos teszteléshez legalább egy, erre a célra használható számítógépekkel berendezett tanteremre van szükség. Ha ezeket a tantermeket a számítógépes tesztelés céljaira kellene létrehozni, az vállalhatatlan beruházást jelentene, és a fejlesztés költségei a PP tesztek alkalmazásával szemben csak sok év után térülnének meg. Egészen más a helyzet, ha ezek a tantermek már ott vannak az iskolában, és többek között erre a célra is fel lehet azokat használni: így beruházás nélkül azonnal jelentkezik a költséghatékonyság előnye. A központi hardver és szoftver felállítása, a feladatbank kifejlesztése a PP tesztek elkészítésénél költségesebb, de karbantartása és alkalmazása már kevésbé költséges.

Az adaptív teszteléshez elegendő iskolánként egy tanteremmel számolni, ahol a párhuzamos osztályok egymás után oldhatják meg a feladatokat. Az adaptív feladatkiosztás biztosítja, hogy a tanulók sokféle feladattal találkoznak, ezért egyrészt nem kell azzal a problémával számolni, hogy a párhuzamos osztályokban tanuló diákok elmondják egymásnak a feladatokat. Az online tesztelés következtében pedig elegendő egy böngészőprogram, aminek segítségével elérhető a központi szerveren futó tesztelőprogram és feladatbank. A szabályosan felszerelt gépekre tehát lényegében semmit nem kell a tesztelés érdekében telepíteni. Ebből a szempontból tehát Magyarországon hamarosan meglesznek az online tesztelés iskolai feltételei, így ezek azok az évek, amikor már fel lehet vetni az online tesztelés elterjesztésének kérdését.

A technikai feltételek megteremtése mellett nehezebb kérdés a társadalmi feltételek megteremtése. Időbe telik, amíg minden érintett (diákok, tanárok, szülők, döntéshozók) megismeri és elfogadja a tesztelés új lehetőségeit. A személyre szabott számítógépes, online tesztelés Amerikában már jelenős múlttal rendelkezik, Európában azonban még csak most kezdődtek meg a szélesebb körű iskolai alkalmazással kapcsolatos kísérletek. Rendkívül fontos, hogy mielőtt bármilyen komoly tétel bíró számítógépes tesztelés elkezdődik, lehetőség legyen a rendszer megismerésére, és az alkalmazás feltételeiről szakmai konszenzus alakuljon ki.

A számítógép-alapú teszteléssel kapcsolatosan az egyik legtöbbet vitatott kérdés a diákok és a tesztelést vezető személy informatikai jártasságának (ICT literacy, ICT familiarity) teszteredményeket befolyásoló hatása, amelyek a kulturális, etnikai és a nemek közötti teljesítménykülönbségek, az emberek között lévő digitális szakadék (digital gap) hatásának felerősödéséhez vezethetnek. Ez a problémakör további validitási kérdéseket is felvethet, mivel ezen a módon az informatikai jártasság vagy a számítógéptől való félelem szintje implicite megjelenik a teszteredményekben is, holott az nem képezte a vizsgálat tárgyát. Az ezen a területen végzett kutatások sem szolgálnak egységes eredménnyel. A kutatási eredmények alapján egyrészről van összefüggés a teszt eredménye és a személy informatikai jártassága között (lásd például: *Tseng, Tiplady és Wright*, 1998), másrészt ez a befolyásoló hatás nem szignifikáns erejű (lásd például: *Powers és O'Neill*, 1993). Általánosabban is megfogalmazhatjuk a kérdést, vajon a tesztelés médiája az informatikai jártasság szintjétől függetlenül bír-e befolyásoló erővel.

Feltehetjük a kérdést, vajon ugyanazt a tudást méri-e a papíralapú és a számítógép-alapú teszt, illetve meddig mérik ugyanazt a tudást. Összehasonlíthatóak-e a különböző médiumon felvett teszteredmények (cross-mode equivalence)? Ezek a kérdések már számos kutatást indukáltak és a mai napig is foglalkoztatják a kutatókat. Az egyes konkrét vizs-

gálatok ugyanis nem adnak még általánosítható választ a problémára. Feltehető, hogy minél inkább megfeleltethető egymásnak flexibilitásban, itemtípusok, alkalmazott elemek tekintetében a papíron, illetve számítógép segítségével kitöltött teszt, annál kisebb a médiahatás. Ezt a feltevést azonban konkrét elemzésekkel kell igazolni, és meg kell határozni, milyen mértékűek az említett hatások. Minél inkább kihasználjuk a számítógép adta lehetőségeket, a számítógép előtt írt és a hagyományos tesztek különböző feladattípusain elért eredmények annál inkább eltérnek egymástól. Ezért az online és papíralapú tesztek eredményeinek összehasonlításakor olyan metrikákat/indexeket kell meghatároznunk, amelyek lehetővé teszik a tesztpontszámok átváltását. *(3)*

Jegyzet

(1) A következőkben az elterjedt angol rövidítéseket fogjuk használni, tekintettel arra, hogy egy szűkebb szakmai kör által használt szakterminológia magyarítása ritkán sikerül.

(2) Elérhető: https://mypisa.acer.edu.au/index.php?option=com_content&task=view&id=66&Itemid=451

(3) A tanulmány a T 046659PSP OTKA kutatási program, az Oktatáselméleti Kutatócsoport és az SZTE MTA Képességkutató Csoport keretében készült. A tanulmány írása idején Molnár Gyöngyvér Bolyai János Kutatási Ösztöndíjban részesült.

Irodalom

Becker, J. (2004): *Computergestütztes Adaptives Testen (CAT) von Angst entwickelt auf der Grundlage der Item Response Theorie (IRT).* Digitális disszertáció. Freie Universität, Berlin.

Butcher, J. N. (1987): *Computerized Psychological Assessment: A Practitioner's Guide.* Basic Books, New York.

Csapó Benő (2000): Tudásszintmérő tesztek. In Falus Iván (szerk.): *A pedagógiai kutatás módszerei.* Műszaki Könyvkiadó, Budapest. 277–316.

Csíkszentmihályi Mihály (1997): *Flow. Az áramlat: a tökéletes élmény pszichológiája.* Akadémiai Kiadó, Budapest.

Frey, A. (2007): Adaptives Testen. In: Moosbrugger, H. – Kelava, A. (szerk.): *Testtheorie und Testkonstruktion.* Springer, Berlin, Heidelberg. Megjelenés alatt.

Griffin, P. (1999): *Item Response Modelling: An introduction to the Rasch Model.* Assessment Research Centre Faculty of Education, The University of Melbourne.

Jurecka, A. – Hartig, J. (2007): Computer- und netzwerkbasiertes Assessment. In Hartig, J. és Klieme, E. (szerk.): *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik.* Bundesministerium für Bildung und Forschung (BMBF), Bonn, Berlin. 37–48.

Kárpáti Andrea (2002): *Informatikai „kereszttanterv" – A számítógéppel segített tanítás és tanulás új paradigmája.* 2007. 09. 25-i megtekintés, www.isze.hu/download/10

Molnár Gyöngyvér (2006): A Rasch-modell alkalmazása a társadalomtudományi kutatásokban. *Iskolakultúra,* 12. 99–113.

Molnár Gyöngyvér (2007): Új ICT eszközök alkalmazása az iskolai gyakorlatban. In Korom Erzsébet (szerk.): *Kihívások a XXI. század iskolájában.* Megjelenés alatt.

Molnár Gyöngyvér (2008): A Rasch-modell kiterjesztése nem dichotóm adatok elemzésére: a rangskálás és parciális kredit modell. *Iskolakultúra,* 1. 66–77.

Nagy József (1972): *A témazáró tudásszintmérés gyakorlati kérdései.* Tankönyvkiadó, Budapest.

Nagy József (1976): *Alsó tagozatos szöveges feladatbank.* JATE, Szeged.

OECD (2007): *PISA– The OECD Programme for International Student Assessment.* http://www.oecd.org/dataoecd/51/27/37474503.pdf

Powers, D. – O'Neill, K. (1993): Inexperienced and anxious computer users: Coping with a computer-administered test of academic skills. *Educational Assessment*, 2. 153–173.

Rasch, G. (1960): Probabilistic models for some intelligence and attainment tests. Danish Institute for Educational Research, Copenhagen.

Rose, M. – Hess, V. – Hörhold, M. – Brähler, E. – Klapp, B. F. (1999): Mobile computergestützte psychometrische Diagnostik. Ökonomische Vorteile und Ergebnisse zur Teststabilität. *Psychotherapie Psychosomatik Medizinische Psychologie*, 49. 202–207.

Tseng, H.-M. – Tiplady, B. – Macleod, H. A. – Wright, P. (1998): Computer anxiety: a comparison of pen-based personal digital assistants, conventional computer, and paper assessment of mood and performance. *British Journal of Psychology*, 89. 599–610.

Van der Linden, W. J. – Ariel, A. – Veldkamp, B. P. (2006): Assembling a Computerised Adaptive Testing Item Pool as a Set of Linear Tests. *Journal of Educational and Behavioral Statistics*, 1. 81–99.

Weiss, D. J. (2004): Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development*, 2. 70–84.